Bernoulli **20**(4), 2014, 1999–2019
DOI: 10.3150/13-BEJ548

Optimal filtering and the dual process

OMIROS PAPASPILIOPOULOS[1] and MATTEO RUGGIERO[2]

[1]*ICREA & Department of Economics and Business, Universitat Pompeu Fabra, Ramón Trias Fargas 25-27, 08005, Barcelona, Spain. E-mail: omiros.papaspiliopoulos@upf.edu*
[2]*Collegio Carlo Alberto & Department of Economics and Statistics, University of Torino, C.so Unione Sovietica 218/bis, 10134, Torino, Italy. E-mail: matteo.ruggiero@unito.it*

We link optimal filtering for hidden Markov models to the notion of duality for Markov processes. We show that when the signal is dual to a process that has two components, one deterministic and one a pure death process, and with respect to functions that define changes of measure conjugate to the emission density, the filtering distributions evolve in the family of finite mixtures of such measures and the filter can be computed at a cost that is polynomial in the number of observations. Special cases of our framework include the Kalman filter, and computable filters for the Cox–Ingersoll–Ross process and the one-dimensional Wright–Fisher process, which have been investigated before. The dual we obtain for the Cox–Ingersoll–Ross process appears to be new in the literature.

Keywords: Bayesian conjugacy; Cox–Ingersoll–Ross process; finite mixture models; hidden Markov model; Kalman filter

1. Introduction

A hidden Markov model (HMM) for a sequence of observations $\{Y_n, n \geq 0\}$, where $Y_n \in \mathcal{Y}$, is a discrete-time stochastic process with dynamics depicted in Figure 1. It is defined in terms of a hidden Markov chain, the so-called *signal*, which in this paper will be taken to be the discrete-time sampling of a time-homogeneous continuous-time Markov process X_t, with state-space $\mathcal{X}$, transition kernel $P_t(x, \mathrm{d}x')$, and initial distribution $\nu(\mathrm{d}x)$. The observations relate to the signal by means of conditional distributions, assumed to be given by the kernel $F(x, \mathrm{d}y)$. We will assume that

$$F(x, \mathrm{d}y) = f_x(y)\mu(\mathrm{d}y) \tag{1}$$

for some measure $\mu(\mathrm{d}y)$, in which case the corresponding densities are known as the *observation* or *emission* densities. The optimal filtering problem is the derivation of the conditional distributions $\mathcal{L}(X_{t_n}|Y_0, \ldots, Y_n)$ of the unobserved signal given the observations collected up to time t_n, henceforth denoted $\nu_n(\mathrm{d}x)$. These filtering distributions are the backbone of all statistical estimation problems in this framework, such as the prediction of future observations, the derivation of smoothing distributions (i.e., the conditional distribution of X_{t_n} given past and future observations) and the calculation of the likelihood function, that is, the marginal density of the observations when the emission distributions are dominated. See [3] for details and applications.

Throughout the paper, we will assume that the signal is stationary and reversible with respect to a probability measure π. Section 4 shows how to extend our result to non-stationary signals. It is also appealing, from a modeling point of view, to assume that the signal evolves in continuous

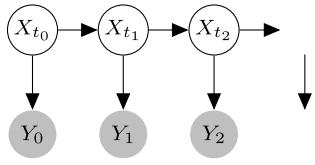

Figure 1. Hidden Markov model represented as a graphical model.

time, since there is a rich family of such models with a prespecified stationary measure π. In addition, this assumption will give us a powerful tool to study optimal filtering by using the generator of the process, as we show in Section 2. In the examples of Section 3, the state space $\mathcal{X}$ of the signal will either be a subset of $\mathbb{R}$ or the $(K-1)$-dimensional simplex Δ_K.

Mathematically, optimal filtering is the solution of the recursion

$$\nu_0 = \phi_{Y_0}(\nu), \qquad \nu_n = \phi_{Y_n}\big(\psi_{t_n - t_{n-1}}(\nu_n)\big), \qquad n > 0,$$

which involves the following two operators acting on probability measures ξ:

$$\begin{aligned} &\text{update:} && \phi_y(\xi)(\mathrm{d}x) = \frac{f_x(y)\xi(\mathrm{d}x)}{p_\xi(y)}, \qquad p_\xi(y) = \int_{\mathcal{X}} f_x(y)\xi(\mathrm{d}x), \\ &\text{prediction:} && \psi_t(\xi)\big(\mathrm{d}x'\big) = \xi P_t\big(\mathrm{d}x'\big) = \int_{\mathcal{X}} \xi(\mathrm{d}x) P_t\big(x, \mathrm{d}x'\big). \end{aligned} \tag{2}$$

The "update" is the application of Bayes theorem, and the "prediction" gives the distribution of the next step of the Markov chain initiated from ξ. These operators have the following property when applied to *finite mixtures of distributions*:

$$\phi_y\left(\sum_{i=1}^n w_i \xi_i\right)(\mathrm{d}x) = \sum_{i=1}^n \frac{w_i\, p_{\xi_i}(y)}{\sum_j w_j\, p_{\xi_j}(y)} \phi_y(\xi_i), \qquad \psi_t\left(\sum_{i=1}^n w_i \xi_i\right)(\mathrm{d}x) = \sum_{i=1}^n w_i \psi_t(\xi_i). \tag{3}$$

This implies that when $\mathcal{X}$ is a finite set, there is a simple algorithm for the sequential computation of the filtering probabilities. To see this, note that we can think of a distribution ν on a finite set $\mathcal{X}$, specified in terms of probabilities $\alpha_x, x \in \mathcal{X}$, as a finite mixture of point masses, $\nu = \sum_x \alpha_x \delta_x$; it is easy to compute $\phi_y(\delta_x)$, $\psi_t(\delta_x)$ and then use the above result to obtain the probabilities associated with the distributions $\phi_y(\nu)$ and $\psi_t(\nu)$. This yields a popular algorithm for inference in HMMs, commonly known as the Baum–Welch filter, whose complexity is easily seen to be $\mathcal{O}(n|\mathcal{X}|^2)$, where $|\mathcal{X}|$ is the cardinality of $\mathcal{X}$.

Outside the finite state-space case, the iteration of these two operators typically leads to analytically intractable distributions. However, there are notable exceptions to this rule. The classic example is the linear Gaussian state-space model, for which the filtering distributions are Gaussian with mean and covariance that can be iteratively computed using the so-called Kalman filter, at cost that grows linearly with n. Recent work by Genon-Catalot and collaborators uncovered that there exist interesting non-Gaussian models for which the filtering distributions are *finite mixtures* of parametric distributions. See [4,5,13], where the authors show how to compute the

corresponding parameters sequentially in these models. We revisit their findings in Section 3. However, the number of mixture components increases with n in a way such that the cost of computing the filters grows polynomially with n (see Section 2 for details). Borrowing and adapting the terminology from [4], we will refer to filters with such computational cost as *computable*, whereas filters whose cost grows linearly with n as *finite-dimensional*.

The work by Genon-Catalot and collaborators raises four important questions, which we address in this paper: are there more models which admit computable filters; do they share some basic structure; is there a general methodology to identify such models and to obtain the algorithm which computes the sequence of parameters; what is the computational complexity of such schemes and how can we obtain faster approximate filtering algorithms? We show that the answer to all these questions relates to an important probabilistic object: the *dual process*. Duality methods have a long history in Probability, dating back to the work of P. Lévy [20] (see [16] for a recent review). These have been widely applied to the study of interacting particle systems [21] and proven to be a powerful method which provides alternative, and often simpler, tools for investigating the sample path properties of the process at hand. For example, the existence of a dual for a certain Markov process (and for a sufficiently large class of functions) implies that the associated martingale problem is well defined, hence that the process is unique; see Section 4.4 of [10]. See also [7] and [8] for applications of duality to population genetics.

In this paper, we illustrate that dual processes play a central role in optimal filtering and to a great extent can be used to settle the four questions posed above. We also uncover their potential as auxiliary variables in Monte Carlo schemes for stochastic processes (and, hence, as a variance reduction scheme). In our framework, the dual will in general be given by two components: a *deterministic process*, driven by an ordinary differential equation, and a (multidimensional) *death process* with countable state-space. We show how to derive an explicit, recursive filtering scheme once the dual is identified, and apply this methodology to three cases of fundamental interest. In doing so, we identify what, to the best of our knowledge, is a new *gamma-type* duality.

The rest of the paper is organized as follows. In Section 2, we link optimal filtering to a specific type of duality, we show how to identify the dual in terms of the *generator* of X_t, and study the complexity of the resulting filtering algorithm. Section 3 analyzes three interesting models for which the dual process is derived: the Cox–Ingersoll–Ross model, the Ornstein–Uhlenbeck process and the K-dimensional Wright–Fisher diffusion. These models are reversible with respect to the gamma, Gaussian and Dirichlet distribution, respectively, and for the Gaussian case the computable filter reduces to the Kalman filter. Section 4 discusses certain aspects of the methodology, including the extension to infinite-dimensional signals modeled as Fleming–Viot processes.

2. Methodology: Filtering the dual process

2.1. Linking optimal filtering to duality

Before presenting the main results, we introduce three fundamental assumptions which provide the general framework under which the results are derived. First, we will assume that X is reversible with respect to a probability measure π:

A1 (Reversibility): $\pi(\mathrm{d}x)P_t(x,\mathrm{d}x') = \pi(\mathrm{d}x')P_t(x',\mathrm{d}x)$.

Section 4 discusses how this assumption can be relaxed to accommodate non-stationary signals. In order to state the second assumption, we need to introduce a certain amount of notation. Define, for $K \in \mathbb{Z}_+ = \mathbb{N} \cup \{0\}$, the space of multi-indices

$$\mathcal{M} = \mathbb{Z}_+^K = \big\{\mathbf{m} = (m_1, \ldots, m_K) : m_j \in \mathbb{Z}_+, j = 1, \ldots, K\big\}. \tag{4}$$

We will use the symbol $\mathbf{0}$ to denote the vector of zeros, $\mathbf{e}_j$ for the vector in $\mathcal{M}$ whose only non-zero element is found at the jth coordinate and equals 1, and let $|\mathbf{m}| = \sum_i m_i$. Furthermore, we will use the product order on $\mathcal{M}$, according to which for $\mathbf{m}, \mathbf{n} \in \mathcal{M}$, $\mathbf{m} \leq \mathbf{n}$ if an only if $m_j \leq n_j$ for all j. Then, for $\mathbf{i} \leq \mathbf{m}$, $\mathbf{m} - \mathbf{i}$ is the vector with jth element $m_j - i_j$. Additionally, if $\Lambda \subset \mathcal{M}$, define

$$G(\Lambda) = \{\mathbf{n} \in \mathcal{M} : \mathbf{n} \leq \mathbf{m}, \mathbf{m} \in \Lambda\}. \tag{5}$$

The notation for $\mathcal{M}$ does not reflect its dependence on the dimension K, but we will reserve boldface for elements of $\mathcal{M}$ when $K > 1$ (or unspecified), whereas normal typeface will be used for elements of $\mathbb{Z}_+$. Finally, the following notations will be used to denote conditional expectations

$$(P_t f)(x) = \mathbb{E}^x\big[f(X_t)\big] = \mathbb{E}\big[f(X_t)|X_0 = x\big] = \int_{\mathcal{X}} f\big(x'\big)P_t\big(x,\mathrm{d}x'\big).$$

The first denotes the action on f of the semigroup operator associated to the transition kernel, where with some abuse of notation the same symbol is used both for the semigroup and the kernel.

The second assumption is concerned with models where $\pi(\mathrm{d}x)$ is *conjugate* to the emission density $f_x(y)$:

A2 (Conjugacy): For $\Theta \subseteq \mathbb{R}^l$, $l \in \mathbb{Z}_+$, let $h : \mathcal{X} \times \mathcal{M} \times \Theta \to \mathbb{R}_+$ be such that $\sup_x h(x, \mathbf{m}, \theta) < \infty$ for all $\mathbf{m} \in \mathcal{M}, \theta \in \Theta$, and $h(x, \mathbf{0}, \tilde{\theta}) = 1$ for some $\tilde{\theta} \in \Theta$. Then $\mathcal{F} = \{h(x, \mathbf{m}, \theta) \times \pi(\mathrm{d}x), \mathbf{m} \in \mathcal{M}, \theta \in \Theta\}$ is assumed to be a family of probability measures such that there exist functions $t : \mathcal{Y} \times \mathcal{M} \to \mathcal{M}$ and $T : \mathcal{Y} \times \Theta \to \Theta$ with $\mathbf{m} \to t(y, \mathbf{m})$ increasing and such that

$$\phi_y\big(h(x, \mathbf{m}, \theta)\pi(\mathrm{d}x)\big) = h\big(x, t(y, \mathbf{m}), T(y, \theta)\big)\pi(\mathrm{d}x).$$

Hence here with conjugacy, we intend the fact that the family $\mathcal{F}$ of measures, which includes π, is closed under the update operation. The assumption that h is bounded in x will be discussed after the statement of Assumption A4.

For $p_\nu(y)$ as in (2), it is easy to check that in the context of A2, we have

$$p_{h(x,\mathbf{m},\theta)\pi(\mathrm{d}x)}(y) =: c(\mathbf{m}, \theta, y) = \frac{f_x(y)h(x, \mathbf{m}, \theta)}{h(x, t(y, m), T(y, \theta))}, \tag{6}$$

which, despite its appearance, does not depend on x.

Note that our definitions of $\mathcal{M}$ and Θ allow the possibility that $K = 0$ or $l = 0$, in which case h in A2 is function only of the variables with non-zero dimension, whereas the case $K = l = 0$

is not of interest here. In the setting of Assumption A2 and for the trivial Markov dynamics $X_t \equiv X_0$, with $X_0 \sim \pi$, the filtering problem collapses to conjugate Bayesian inference for the unknown parameter x of the sampling density $f_x(y)$. See Section 5.2 and Appendix A.2 of [2] for an exposition of conjugate Bayesian inference and stylized conjugate Bayesian models, and Section 3 in this paper for examples within our framework.

The third main assumption for our results concerns the existence of a certain type of dual process for the signal.

A3 (Duality): We assume that $r : \Theta \to \Theta$ is such that the differential equation

$$\mathrm{d}\Theta_t/\mathrm{d}t = r(\Theta_t), \qquad \Theta_0 = \theta_0, \tag{7}$$

has a unique solution for all θ_0. Let $\lambda : \mathbb{Z}_+ \to \mathbb{R}_+$ be an increasing function, $\rho : \Theta \to \mathbb{R}_+$ be a continuous function, and consider a two-component Markov process (M_t, Θ_t) with state-space $\mathcal{M} \times \Theta$, where Θ_t evolves autonomously according to (7), and when at $(M_t, \Theta_t) = (\mathbf{m}, \theta)$, the process jumps down to state $(\mathbf{m} - \mathbf{e}_j, \theta)$ with instantaneous rate

$$\lambda\big(|\mathbf{m}|\big)\rho(\theta)m_j. \tag{8}$$

We assume (M_t, Θ_t) is *dual* to X_t with respect to the family of functions h defined in A2, in the sense that

$$\mathbb{E}^x\big[h(X_t, \mathbf{m}, \theta)\big] = \mathbb{E}^{(\mathbf{m},\theta)}\big[h(x, M_t, \Theta_t)\big] \qquad \forall x \in \mathcal{X}, \mathbf{m} \in \mathcal{M}, \theta \in \Theta, t \geq 0. \tag{9}$$

When $K = 0$ or $l = 0$ in A2, the dual process is just Θ_t or M_t, respectively, and we adopt the convention that

$$\rho(\theta) \equiv 1 \quad \text{whenever} \quad l = 0.$$

Note that M_t can only jump to "smaller" states according to the partial order on $\mathcal{M}$, and that (8) implies that 0 is an absorbing state for each coordinate j of M_t, so that the vector of zeros is a global absorbing state.

As mentioned in Section 1, the notion of duality for Markov processes with respect to a given function is well known. See, for example, Section II.4 in [21]. Among the most common type of duality relations we mention *moment duality*, that is duality with respect to functions of type $h(x, y) = x^y$, and *Laplace duality*, that is with respect to functions of type $h(x, y) = \mathrm{e}^{-axy}$. See, for example, [16]. In our framework, the duality functions are Radon–Nikodym derivatives between measures that are conjugate to the emission density, and this setup is perfectly tailored to optimal filtering. Furthermore, A3 specifies that we are interested in dual processes which can be decomposed into two parts: one purely *deterministic* and the other given by a K-dimensional *pure death process*, whose death rates are subordinated by the deterministic process. The transition probabilities of the death process, conditional on the initial state $\Theta_0 = \theta$, will be denoted by

$$p_{\mathbf{m},\mathbf{n}}(t; \theta) = \mathbb{P}[M_t = \mathbf{n} | M_0 = \mathbf{m}, \Theta_0 = \theta], \qquad \mathbf{n}, \mathbf{m} \in \mathcal{M}, \mathbf{n} \leq \mathbf{m}. \tag{10}$$

It is worth mentioning that the requirements on the structure of the dual processes prescribed by Assumption A3, with particular reference to the intensity (8), are justified by the three main rea-

sons. The first is that, as shown in Section 3, they define a framework general enough to identify duals of processes of interest, the incorporation of a deterministic component being necessary in this respect. The second reason is that the transition probabilities (10) are analytically available, as provided by the following result, whose proof can be found in the Appendix.

Proposition 2.1. *Let* (M_t, Θ_t) *be as in* A3, *with* $(M_0, \Theta_0) = (\mathbf{m}, \theta) \in \mathcal{M} \times \Theta$, *and let* $\lambda_{|\mathbf{m}|} = |\mathbf{m}|\lambda(|\mathbf{m}|)$. *Then the transition probabilities for* M_t *are* $p_{\mathbf{m},\mathbf{m}}(t;\theta) = \exp\{-\lambda_{|\mathbf{m}|}\int_0^t \rho(\Theta_s)\,\mathrm{d}s\}$ *and, for any* $\mathbf{0} \le \mathbf{i} \le \mathbf{m}$,

$$p_{\mathbf{m},\mathbf{m}-\mathbf{i}}(t;\theta) = \left(\prod_{h=0}^{|\mathbf{i}|-1} \lambda_{|\mathbf{m}|-h}\right) C_{|\mathbf{m}|,|\mathbf{m}|-|\mathbf{i}|}(t)\, p\big(i_1,\ldots,i_K;\mathbf{m},|\mathbf{i}|\big),$$

where

$$C_{|\mathbf{m}|,|\mathbf{m}|-|\mathbf{i}|}(t) = (-1)^{|\mathbf{i}|} \sum_{k=0}^{|\mathbf{i}|} \frac{\mathrm{e}^{-\lambda_{|\mathbf{m}|-k}\int_0^t \rho(\Theta_s)\,\mathrm{d}s}}{\prod_{0\le h\le|\mathbf{i}|,h\ne k}(\lambda_{|\mathbf{m}|-k} - \lambda_{|\mathbf{m}|-h})}$$

and $p(i_1,\ldots,i_K;\mathbf{m},|\mathbf{i}|)$ *is the multivariate hypergeometric probability mass function with parameters* $(\mathbf{m},|\mathbf{i}|)$ *evaluated at* $(i_1,\ldots,i_K)$.

This result can be interpreted as follows. The probability that a one-dimensional death process with inhomogeneous rates $\lambda_{|\mathbf{m}|}\rho(\Theta_s)$ decreases from $|\mathbf{m}|$ to $|\mathbf{m}| - |\mathbf{i}|$ in the interval $[0,t]$ is $(\prod_{h=0}^{|\mathbf{i}|-1}\lambda_{|\mathbf{m}|-h})C_{|\mathbf{m}|,|\mathbf{m}|-|\mathbf{i}|}(t)$, where the second factor is related to the convolution of the waiting times in an inhomogeneous Poisson process (see Section 19.10 in [17], and [22]). For a K-dimensional death process, the same quantity is the probability associated to all paths leading from level $|\mathbf{m}|$ to level $|\mathbf{m}| - |\mathbf{i}|$. Given such event, the probability of the subset of paths leading exactly from $\mathbf{m}$ to $\mathbf{m} - \mathbf{i}$ is then given by the multivariate hypergeometric probability $p(i_1,\ldots,i_K;\mathbf{m},|\mathbf{i}|)$ (an expression of this probability can be found in the Appendix).

Note that the special case of Proposition 2.1 yielded by $K = 1$ and $\rho(\Theta_s) \equiv 1$ relates to the result obtained in Proposition 4.5 in [5]. Note also that when $\rho(\Theta_s) \equiv 1$ and $\lambda_m = m(\theta + m - 1)/2$, $C_{|\mathbf{m}|,|\mathbf{m}|-|\mathbf{i}|}(t)$ is the transition probability of the block-counting process of Kingman's coalescent with mutation, see [23] and [14] for details on such process.

The third motivation behind the type of duality required by A3 is that if it holds, the prediction operator maps measures as in A2 into finite mixtures.

Proposition 2.2. *Let* ψ_t *be as in* (2) *and assume* A1–A2–A3 *hold. Then*

$$\psi_t\big(h(x,\mathbf{m},\theta)\pi(\mathrm{d}x)\big) = \sum_{\mathbf{0}\le\mathbf{i}\le\mathbf{m}} p_{\mathbf{m},\mathbf{m}-\mathbf{i}}(t;\theta)h(x,\mathbf{m}-\mathbf{i},\Theta_t)\pi(\mathrm{d}x) \tag{11}$$

with $p_{\mathbf{m},\mathbf{m}-\mathbf{i}}(t;\theta)$ *as in Proposition* 2.1 *and where* Θ_t *is the value in* t *of the process in* (7) *started from* $\Theta_0 = \theta$.

Proof. From (2), we have

$$\begin{aligned}\psi_t\big(h(x,\mathbf{m},\theta)\pi(\mathrm{d}x)\big) &= \int_{\mathcal{X}} h(x,\mathbf{m},\theta)\pi(\mathrm{d}x)\,P_t\big(x,\mathrm{d}x'\big) = \int_{\mathcal{X}} h(x,\mathbf{m},\theta)\pi\big(\mathrm{d}x'\big)\,P_t\big(x',\mathrm{d}x\big)\\ &= \pi\big(\mathrm{d}x'\big)\mathbb{E}^{x'}\big[h(X_t,\mathbf{m},\theta)\big] = \pi\big(\mathrm{d}x'\big)\mathbb{E}^{(\mathbf{m},\theta)}\big[h\big(x',M_t,\Theta_t\big)\big]\\ &= \sum_{\mathbf{n}\le\mathbf{m}} p_{\mathbf{m},\mathbf{n}}(t;\theta)h\big(x',\mathbf{n},\Theta_t\big)\pi\big(\mathrm{d}x'\big),\end{aligned}$$

where the second equality follows from A1, the fourth from A3, and the last from (10). □

The above result states that reversibility and the existence of the required duality jointly guarantee that the prediction operator can be computed with a finite effort. The reduction of the operator to a sum is due to the fact that X_t is dual to a Markov process with discrete state-space, but it is precisely the fact that M_t is a pure death process that makes the number of terms in the sum being finite. The next result shows that computable filtering is available in the framework we have outlined.

Proposition 2.3. *Consider the family of finite mixtures*

$$\bar{\mathcal{F}}_f = \left\{\sum_{\mathbf{m}\in\Lambda} w_{\mathbf{m}}h(x,\mathbf{m},\theta)\pi(\mathrm{d}x) : \Lambda\subset\mathcal{M}, |\Lambda|<\infty, w_{\mathbf{m}}\ge 0, \sum_{\mathbf{m}\in\Lambda} w_{\mathbf{m}} = 1\right\}. \tag{12}$$

Then, under Assumptions A1–A2–A3, $\bar{\mathcal{F}}_f$ *is closed under the application of the prediction and update operators* (2), *and specifically*

$$\phi_y\left(\sum_{\mathbf{m}\in\Lambda} w_{\mathbf{m}}h(x,\mathbf{m},\theta)\pi(\mathrm{d}x)\right) = \sum_{\mathbf{n}\in t(y,\Lambda)} \widehat{w}_{\mathbf{n}}h\big(x,\mathbf{n},T(y,\theta)\big)\pi(\mathrm{d}x)$$

with

$$\begin{aligned} t(y,\Lambda) &:= \big\{\mathbf{n} : \mathbf{n} = t(y,\mathbf{m}), \mathbf{m}\in\Lambda\big\}\\ \widehat{w}_{\mathbf{n}} &\propto w_{\mathbf{m}}c(\mathbf{m},\theta,y) \qquad \textit{for } \mathbf{n} = t(y,\mathbf{m}), \sum_{\mathbf{n}\in t(y,\Lambda)} \widehat{w}_{\mathbf{n}} = 1,\end{aligned} \tag{13}$$

and

$$\psi_t\left(\sum_{\mathbf{m}\in\Lambda} w_{\mathbf{m}}h(x,\mathbf{m},\theta)\pi(\mathrm{d}x)\right) = \sum_{\mathbf{n}\in G(\Lambda)}\left(\sum_{\mathbf{m}\in\Lambda,\mathbf{m}\ge\mathbf{n}} w_{\mathbf{m}}p_{\mathbf{m},\mathbf{n}}(t;\theta)\right)h(x,\mathbf{n},\theta_t)\pi(\mathrm{d}x). \tag{14}$$

The above proposition shows that under Assumption A1 to A3, and provided the starting state belongs to the family $\bar{\mathcal{F}}_f$ of finite mixtures with components as in A2, then the filtering distributions evolve within $\bar{\mathcal{F}}_f$. Furthermore, the explicit reweighing of the mixture components is provided, thus allowing to concretely implement the recursive filtering scheme. Note also that

this result generalizes Theorem 2.1 in [4], which states a similar result for $K = 1$ under the Assumption A2 and the result in Proposition 2.2. The proof of Proposition 2.3 follows from (3), A2 and Proposition 2.2 by direct computation, and is thus omitted. Later in this section, we will derive filtering algorithms based on this result. However, we first address in the next subsection the most important aspect of the approach described in this section, which is how to find a dual process that satisfies A2.

2.2. Local duality as a sufficient condition

It is typically easier to identify a process that satisfies the duality relation (9) for infinitesimal t. Formally, this requires studying the *generator* of X_t, which we will denote by $\mathcal{A}$. This is a linear operator, with domain denoted $\mathscr{D}(\mathcal{A})$, linked to the semigroup operator via the Kolmogorov backward equation

$$\frac{\partial}{\partial t} P_t f(x) = (\mathcal{A} P_t f)(x), \qquad f \in \mathscr{D}(\mathcal{A}),$$

where on the left hand side $P_t h(x)$ is differentiated in t for given x, whereas on the right hand side, $\mathcal{A}$ acts on $P_t h(x)$ as a function of x for given t. See, for example, Proposition 1.1.5 in [10].

Suppose now X_t is a diffusion process which solves an SDE on $\mathbb{R}^d$ of the form

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}B_t.$$

In this case, $\mathcal{A}$ is the second-order differential operator given by

$$(\mathcal{A} f)(x) = \sum_{i=1}^{d} b_i(x) \frac{\partial f(x)}{\partial x_i} + \frac{1}{2} \sum_{i,j=1}^{d} a_{i,j}(x) \frac{\partial^2 f(x)}{\partial x_i\, \partial x_j}, \qquad f \in \mathscr{D}(\mathcal{A}), \tag{15}$$

for an appropriate domain $\mathscr{D}(\mathcal{A})$ and where $a_{i,j}(x) := (\sigma(x)\sigma(x)^{\mathrm{T}})_{i,j}$.

Let now A denote the generator of the dual process defined in A3, which can be easily checked to be

$$\begin{aligned}(Ag)(\mathbf{m}, \theta) = \lambda\big(|\mathbf{m}|\big)\rho(\theta) \sum_{i=1}^{K} m_i \big[g(\mathbf{m} - \mathbf{e}_i, \theta) - g(\mathbf{m}, \theta)\big] \\ + \sum_{i=1}^{l} r_i(\theta) \frac{\partial g(\mathbf{m}, \theta)}{\partial \theta}, \qquad g \in \mathscr{D}(A),\end{aligned} \tag{16}$$

with r as in (7). The main idea is then to identify the dual process from the generator, instead of the semigroup operator.

A4 (Local duality): The function $h(x, \mathbf{m}, \theta)$ defined in A2 is such that $h(x, \mathbf{m}, \theta)$, as a function of x belongs to $\mathscr{D}(\mathcal{A})$ for all $(\mathbf{m}, \theta) \in \mathcal{M} \times \Theta$, as a function of $(\mathbf{m}, \theta)$ belongs to $\mathscr{D}(A)$ for all $x \in \mathcal{X}$, and

$$\big(\mathcal{A} h(\cdot, \mathbf{m}, \theta)\big)(x) = \big(A h(x, \cdot, \cdot)\big)(\mathbf{m}, \theta) \qquad \forall x \in \mathcal{X}, \mathbf{m} \in \mathcal{M}, \theta \in \Theta. \tag{17}$$

It will typically be the case that in order to have $h(\cdot, \mathbf{m}, \theta) \in \mathscr{D}(\mathcal{A})$, one needs h to be bounded in x, requirement included in Assumption A2. Proposition 1.2 of [16] shows that A4, together with the further assumption

$$\mathbb{E}^x\big[h(X_t, \mathbf{m}, \theta)\big] \in \mathscr{D}(A), \qquad \mathbb{E}^{(\mathbf{m},\theta)}\big[h(x, M_t, \Theta_t)\big] \in \mathscr{D}(\mathcal{A}),$$

implies (9), whose argument can be sketched as follows. From (17), we can write

$$(\beta I - \mathcal{A})h = (\beta I - A)h, \qquad \beta \in \mathbb{R},$$

where I denotes the identity operator. Since $\mathcal{A}$ and A generate strongly continuous contraction semigroups, say on L_1 and L_2, their ranges are dense in L_1 and L_2, respectively. Moreover, the resolvents $\mathcal{R}_\beta = (\beta I - \mathcal{A})^{-1}$, $R_\beta = (\beta I - A)^{-1}$ are one-to-one for all $\beta > 0$, so the previous implies

$$\mathcal{R}_\beta h = R_\beta h, \qquad h \in L_1 \cap L_2, \beta > 0.$$

Since the resolvent of an operator is the Laplace transform of the associated semigroup, and because of the uniqueness of Laplace transforms, the previous expression in turn implies (9).

The approach sketched above for identifying the dual process by means of the local condition (17) will be implemented in Section 3, where we will identify the duals for some interesting relevant models.

2.3. The filtering algorithm

Typically, the initial distribution of the signal process belongs to $\mathcal{F}$, and most often equals the invariant measure π. Thus, without loss of generality and in order to simplify the exposition below, we make the following additional assumption.

A5 (Initialization): The initial distribution of the signal is $\nu = h(x, \mathbf{m}_0, \theta_0)\pi(\mathrm{d}x) \in \mathcal{F}$, for some $\mathbf{m}_0 \in \mathcal{M}$, $\theta_0 \in \Theta$.

Proposition 2.3 provides a probabilistic interpretation of the weights involved in the finite mixtures in terms of the transition probabilities of the dual death process M_t. This interpretation can be elaborated further, in order to facilitate the development of filtering algorithms. With a little abuse of notation, denote by $\{D_n = (M_n, \Theta_n), n \geq 0\}$ a discrete-time process with state-space $\mathcal{M} \times \Theta$ constructed as follows. Consider a partially observed Markov process, where the signal is now D_n and the conditional independence structure, given in Figure 2 graphically, is as follows. Let $D_0 = (M_0, \Theta_0) = (\mathbf{m}_0, \theta_0)$ be the initial state of the chain, with $(\mathbf{m}_0, \theta_0)$ defined in A5. Then $\mathcal{L}(Y_n | D_n = (\mathbf{m}, \theta)) = c(\mathbf{m}, \theta, y)\mu(\mathrm{d}y)$, with $c(\mathbf{m}, \theta, y)$ as in (6) and μ in (1), and for $n \geq 1$, $\mathcal{L}(D_n | Y_{n-1} = y, D_{n-1} = (\mathbf{m}, \theta))$ is the law of $(M_{t_n - t_{n-1}}, \Theta_{t_n - t_{n-1}})$ in A3 started from $(t(y, \mathbf{m}), T(y, \theta))$ at time 0. Then, the connection between duality and optimal filtering can be expressed as

$$\mathcal{L}(X_{t_n} | Y_0, \ldots, Y_n) = \int h\big(x, t(Y_n, M_n), T(Y_n, \Theta_n)\big)\pi(\mathrm{d}x)\, \mathrm{d}\mathcal{L}(D_n | Y_0, \ldots, Y_{n-1}). \tag{18}$$

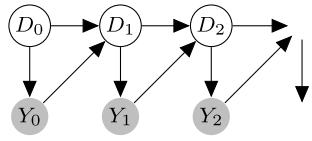

Figure 2. The partially observed Markov process dual to the hidden Markov model in Figure 1, where $D_i = (M_i, \Theta_i)$.

Thus, filtering X_{t_i} in the original model in Figure 1 can be achieved by filtering D_i in the dual model in Figure 2. Since Θ_n evolves deterministically, optimal filtering for X reduces to filtering M_n, which has finite support with probabilities that can be computed recursively using an algorithm similar to the Baum–Welch filter, as we now describe. $\mathcal{L}(M_0, \Theta_0)$ has support on the single point $\{(\mathbf{m}_0, \theta_0)\}$; if $\mathcal{L}(M_n, \Theta_n | Y_0, \ldots, Y_{n-1})$ has support on $\Lambda_n \times \{\theta_n\}$ for $\Lambda_n \subset \mathcal{M}$ and $\theta_n \in \Theta$, and assigns probability $w_{\mathbf{m}}$ to state $(\mathbf{m}, \theta_n)$, then $\mathcal{L}(M_{n+1}, \Theta_{n+1} | Y_0, \ldots, Y_n)$ has support on $\Lambda_{n+1} \times \{\theta_{n+1}\}$, where $\Lambda_{n+1} = G(t(Y_n, \Lambda_n))$, for G and $t(y, \cdot)$ defined in (5) and (13), respectively, θ_{n+1} the solution of (7) at time $t_{n+1} - t_n$ started from $\Theta_0 = \theta_n$, and the probability associated to state $(\mathbf{n}, \theta_{n+1}) \in \Lambda_{n+1} \times \{\theta_{n+1}\}$ is

$$\mathbb{P}[M_{n+1} = \mathbf{n}, \Theta_{n+1} = \theta_{n+1} | Y_0, \ldots, Y_n] = \sum_{\substack{\mathbf{m} \in \Lambda_n \\ t(Y_n, \mathbf{m}) \geq \mathbf{n}}} w_{\mathbf{m}} p_{t(Y_n, \mathbf{m}), \mathbf{n}}(t_{n+1} - t_n; \theta_n). \tag{19}$$

Therefore, the optimal filtering reduces to the sequential computation of the parameters θ_n, the supports Λ_n and the probabilities on each support point in Λ_n, for $n = 0, 1, \ldots$.

The computation of the probabilities (19) for all $(\mathbf{n}, \theta_{n+1}) \in \Lambda_{n+1} \times \{\theta_{n+1}\}$ can be done at a cost that is at most of order $|\Lambda_{n+1}|^2$. Therefore, the overall cost of computing the filters up to the nth observation is bounded from above by $\sum_{i=0}^{n} |\Lambda_i|^2$. If $|\Lambda_i|$ were constant with i, we would recover the complexity of the Baum–Welch filter, discussed in Section 1. However, $|\Lambda_i|$ increases with i, as a result of the successive operation of G and $t(y, \cdot)$ defined in (5) and (13), respectively. Clearly, it is hard to make further analysis on the computational complexity without some information on $t(y, \cdot)$. Here we will assume that $t(y, \mathbf{m}) = \mathbf{m} + N(y)$, where $N : \mathcal{Y} \to \mathcal{M}$, a structure that is found in all the examples we study in this paper. We then have the following key result. The proof of the lemma is omitted.

Lemma 2.4. *For any $\Lambda \subset \mathcal{M}$ and $\mathbf{m} \in \mathcal{M}$, we define $\Lambda + \mathbf{m} = \{\mathbf{n} + \mathbf{m}; \mathbf{n} \in \Lambda\}$. Then*

$$G\big(G(\Lambda) + \mathbf{m}\big) = G(\Lambda + \mathbf{m}).$$

Proposition 2.5. *Under the assumption that $t(y, \mathbf{m}) = \mathbf{m} + N(y)$, where $N : \mathcal{Y} \to \mathcal{M}$, we have that*

$$|\Lambda_n| = G\left(\mathbf{m}_0 + \sum_{i=1}^{n} N(Y_i)\right) \leq \left(1 + \frac{d_n}{K}\right)^K,$$

where $d_n = |\mathbf{m}_0 + \sum_{i=1}^{n} N(Y_i)|$.

Proof. The equality follows by successive application of Lemma 2.4. For the inequality, notice that $\mathbf{n} := \mathbf{m}_0 + \sum_{i=1}^{n} N(Y_i) \in \mathcal{M}$, with $|\mathbf{n}| = d_n$. Then, by (5), $|G(\mathbf{n})| = \prod_{i=1}^{K}(n_i + 1)$. Then, apply Jensen's inequality to $\log|G(\mathbf{n})|$ to obtain the result. □

When the observations follow a stationary process, d_n will be of order n. Therefore, the complexity of carrying out the computations involved in the filtering recursions up to iteration n, will be $\mathcal{O}(n^{2K})$, where the constant depends on K but not n. We return to the issue of complexity in Section 4.

3. The dual of some stochastic processes

Following the local duality approach outlined in the previous section, here we identify the dual processes for the Cox–Ingersoll–Ross model, diffusion processes with linear coefficients and K-dimensional Wright–Fisher diffusions. In accordance with the rest of the article, we focus on stationary parametrizations of the processes and discuss the non-stationary case in Section 4.

Recall that d, K, l denote the state space dimension for X_t, M_t and Θ_t, respectively.

3.1. CIR processes

The so-called Cox–Ingersoll–Ross (CIR) model is a non-negative one-dimensional diffusion, that solves the SDE

$$\mathrm{d}X_t = \big(\delta\sigma^2 - 2\gamma X_t\big)\,\mathrm{d}t + 2\sigma\sqrt{X_t}\,\mathrm{d}B_t.$$

This name is due to [6] who introduced the model in mathematical finance, although this model had been studied long before in the literature, see, for example, the population growth model in Section 13.C of [18] and the process described in Section 5 of [12]. From a broader perspective, the CIR model can also be seen as a special case of a continuous-state branching process with immigration [19].

The generator of the CIR process is

$$\mathcal{A} = \big(\delta\sigma^2 - 2\gamma x\big)\frac{\mathrm{d}}{\mathrm{d}x} + 2\sigma^2 x\frac{\mathrm{d}^2}{\mathrm{d}x^2}, \qquad \delta, \gamma, \sigma > 0, \tag{20}$$

with domain defined as follows. With the above parametrization, and using Feller's terminology, the boundary point $+\infty$ is natural for all choices of parameters, while 0 is regular if $\delta < 2$ and entrance if $\delta \geq 2$. Define

$$\mathscr{D}_0(\mathcal{A}) = \big\{ f \in C_0\big([0,\infty)\big) \cap C^2\big((0,\infty)\big) : \mathcal{A}f \in C_0\big([0,\infty)\big)\big\},$$

where $C_0([0,\infty))$ is the space of continuous functions vanishing at infinity, and

$$\mathscr{D}(\mathcal{A}) = \begin{cases} f \in \mathscr{D}_0(\mathcal{A}), & \text{if } \delta \geq 2, \\ f \in \mathscr{D}_0(\mathcal{A}) : \lim\limits_{x\to 0} x^{\delta/2} f'(x) = 0, & \text{if } 0 < \delta < 2. \end{cases}$$

Then $\{(f, \mathcal{A}f) : f \in \mathscr{D}(\mathcal{A})\}$ generates a Feller semigroup on $C_0([0,\infty))$. Such choice of the domain for the case $0 < \delta < 2$ guarantees that the boundary 0 is instantaneously reflecting. See Theorems 8.1.2 and 8.2.1 in [10]. In this case, the CIR process is reversible with respect to the gamma distribution

$$\pi \equiv \mathrm{Ga}\big(\delta/2, \gamma/\sigma^2\big).$$

Previous results on duality for the CIR model include a Laplace duality, that is with respect to a function of type $h(x, y) = \mathrm{e}^{-axy}$. See, for example, [15]. Here however we identify a new, gamma-type duality relation, which has as special cases a moment and a Laplace duality. Now, let $d = 1$, $K = 1$ and $l = 1$, and define, for $\theta > 0$, the function

$$h(x, m, \theta) = \frac{\Gamma(\delta/2)}{\Gamma(\delta/2 + m)} \left(\frac{\gamma}{\sigma^2}\right)^{-\delta/2} \theta^{\delta/2+m} x^m \exp\big\{-\big(\theta - \gamma/\sigma^2\big)x\big\}.$$

This function can be identified as the Radon–Nikodym derivative of a $\mathrm{Ga}(\delta/2+m, \theta)$ distribution with respect to π. The family of gamma distributions that arises by varying $m \in \mathbb{Z}_+$ and $\theta > 0$, defines a subset of the family of gamma distributions that is conjugate to emission densities that as a function of x are proportional to

$$x^n \mathrm{e}^{-\lambda x}, \qquad n \in \mathbb{Z}_+, \lambda > 0,$$

in which case t and T in A2 coincide with

$$t(y, m) = n + m, \qquad T(y, \theta) = \theta + \lambda.$$

Such type of emission density arises, for example, for observations $Y_n = n$ distributed as Poisson with intensity λX_{t_n}, giving rise to a dynamic version of the Poisson-gamma conjugate Bayesian model.

On the other hand, $h(\cdot, m, \theta)$ belongs to the domain of $\mathcal{A}$ only when $\theta \geq \gamma/\sigma^2$, in which case $h \in C_0^2([0,\infty))$. In order to be able to use local duality as in A4 we will assume that the family is defined as

$$\mathcal{F} = \big\{h(x, m, \theta)\pi(\mathrm{d}x), m \in \mathbb{Z}_+, \theta \geq \gamma/\sigma^2\big\}$$

but we will return to the case $\theta < \gamma/\sigma^2$ at the end of this subsection. Then a simple computation yields

$$\begin{aligned}\mathcal{A}h(\cdot, m, \theta)(x) = {} & 2m\sigma^2\theta h(x, m-1, \theta) + \sigma^2(\delta + 2m)\big(\theta - \gamma/\sigma^2\big)h(x, m+1, \theta) \\ & - \sigma^2\big[2m\theta + (\delta + 2m)\big(\theta - \gamma/\sigma^2\big)\big]h(x, m, \theta).\end{aligned}$$

Motivated by this structure, and with view to achieving the local duality in (17) we consider a two-component process (M_t, Θ_t) with generator A as in (16), where

$$\lambda(m) = 2\sigma^2, \qquad r(\theta) = 2\sigma^2\theta\big(\gamma/\sigma^2 - \theta\big), \qquad \rho(\theta) = \theta.$$

It is then easy to check that local duality holds, namely

$$\mathcal{A}h(\cdot, m, \theta)(x) = Ah(x, \cdot, \cdot)(m, \theta).$$

Additionally, the conditions that are required to derive (9) from this local duality are satisfied. In this example, the solution of the dynamical system (7) for $\Theta_0 = \theta$ is given by

$$\Theta_t = \frac{\gamma}{\sigma^2} \frac{\theta \mathrm{e}^{2\gamma t}}{\theta \mathrm{e}^{2\gamma t} + \gamma/\sigma^2 - \theta},$$

which in conjunction with Proposition 2.1 implies that the transition probabilities for the death process simplify to binomial probabilities

$$p_{m,m-i}(t;\theta) = \mathrm{Bin}\left(m - i; m, \frac{\gamma}{\sigma^2}\left(\theta \mathrm{e}^{2\gamma t} + \gamma/\sigma^2 - \theta\right)^{-1}\right).$$

Therefore, we have all the ingredients necessary to implement the filtering algorithm. Finally, note that if $\theta_0 \geq \gamma/\sigma^2$, then $\theta_n \geq \gamma/\sigma^2$ for all n.

Notice that the result on the transition probabilities above, together with Proposition 2.2, implies the following interesting property of the CIR process:

$$\begin{aligned} &\psi_t\big(\mathrm{Ga}(m + \delta/2, \theta)\big) \\ &\quad = \sum_{k=0}^{m} \mathrm{Bin}\left(k; m, \frac{\gamma}{\sigma^2}\left(\theta \mathrm{e}^{2\gamma t} + \gamma/\sigma^2 - \theta\right)^{-1}\right) \mathrm{Ga}\left(k + \delta/2, \frac{\gamma}{\sigma^2} \frac{\theta \mathrm{e}^{2\gamma t}}{\theta \mathrm{e}^{2\gamma t} + \gamma/\sigma^2 - \theta}\right). \end{aligned} \tag{21}$$

This result has been obtained before, using a completely different approach; the case $\delta = 1$ can be shown directly by elementary calculations using a change of variables and binomial expansion of the left-hand-side; the general case was proved in [4], see Proposition 3.4 and the associated Lemma 3.1, after some rather heavy calculations. The result in (21) leads to a computable filter, as we showed in Proposition 2.3, which is precisely the result also obtained in [4] for the CIR process. It is neat that using duality and the generic result in Proposition 2.2, this result can be obtained in a straightforward manner. The proof in [4] is based on the following known series expansion of the CIR transition kernel, see expression (80) in [4] and page 334 of [18], which can be re-expressed as a Poisson mixture of gamma distributions as follows:

$$P_t\big(x, \mathrm{d}x'\big) = \sum_{k \geq 0} \mathrm{Poisson}\left(k; \frac{\gamma}{\sigma^2} \frac{1}{\mathrm{e}^{2\gamma t} - 1} x\right) \mathrm{Ga}\left(k + \delta/2, \frac{\gamma}{\sigma^2} \frac{\mathrm{e}^{2\gamma t}}{\mathrm{e}^{2\gamma t} - 1}\right). \tag{22}$$

It is interesting that instead of deriving (21) from (22), which in any case is laborious, one can prove the former using duality and then obtain (22) by taking $\theta = (m + \delta/2)/x$ and letting $m \to \infty$ in (21).

In view of the arguments of Section 2, it follows that, for $\theta < \gamma/\sigma^2$, $h \notin \mathscr{D}(\mathcal{A})$, hence duality in the sense of A3 cannot be established using local duality. However, in view of the result (21) that has already been obtained in [4], it is obvious that duality still holds in this case. This

also shows the limitation of the functional analytic method for establishing duality: it is a very powerful when all formal requirements are met, but there will be examples, like this one, where (9) would have to be established by alternative arguments. Nevertheless, a formal calculation using the generator reveals the dual even when $\theta < \gamma/\sigma^2$.

3.2. Linear diffusion processes

We consider the scalar Ornstein–Uhlenbeck process that solves an SDE of the form

$$\mathrm{d}X_t = -\frac{\sigma^2}{\alpha}(X_t - \gamma)\,\mathrm{d}t + \sqrt{2}\sigma\,\mathrm{d}B_t,$$

which is reversible with respect to the Gaussian distribution,

$$\pi(\mathrm{d}x) \equiv \mathrm{Normal}(\gamma, \alpha).$$

The generator is given by

$$\mathcal{A} = \big(\sigma^2\gamma/\alpha - \sigma^2 x/\alpha\big)\frac{\mathrm{d}}{\mathrm{d}x} + \sigma^2\frac{\mathrm{d}^2}{\mathrm{d}x^2}$$

with domain $C_0^2((-\infty, \infty))$. In this model, we have $d = 1, K = 0, l = 2$, where $\theta = (\mu, \tau) \in \mathbb{R} \times \mathbb{R}_+$, and

$$h(x, \mu, \tau) = \left(\frac{\alpha}{\tau}\right)^{1/2} \exp\left\{-\frac{(x-\mu)^2}{2\tau} + \frac{(x-\gamma)^2}{2\alpha}\right\},$$

which can be easily recognised as the Radon–Nikodym derivative between a $\mathrm{Normal}(\mu, \tau)$ and π. The measures $h(x, \mu, \tau)\pi(\mathrm{d}x)$ are conjugate to emission densities that as a function of x are proportional to

$$\exp\left\{-\frac{1}{2\lambda}(x-c)^2\right\}, \qquad \lambda > 0, c \in \mathbb{R},$$

with $T(y, \theta) = ((\lambda\mu + \tau c), \lambda\tau)/(\lambda + \tau)$. Such density arises, for example, with data $Y_n = c$ that is Gaussian with mean X_{t_n} and variance λ. As with the CIR process, we have the technical problem that this function belongs to $\mathscr{D}(\mathcal{A})$ only for $\tau < \alpha$, hence we will restrict to this case and define $\Theta = \{(\mu, \tau) : \mu \in \mathbb{R}, 0 < \tau < \alpha\}$. A direct calculation gives that

$$\mathcal{A}h(\cdot, \mu, \tau)(x) = \frac{\sigma^2}{\alpha}(\gamma - \mu)\frac{\partial}{\partial\mu}h(x, \mu, \tau) + 2\sigma^2(1 - \tau/\alpha)\frac{\partial}{\partial\tau}h(x, \mu, \tau).$$

This suggests that the dual is purely deterministic and described in terms of the ODEs:

$$\mathrm{d}\mu_t/\mathrm{d}t = \frac{\sigma^2}{\alpha}(\gamma - \mu_t)\,\mathrm{d}t, \qquad \mathrm{d}\tau_t/\mathrm{d}t = 2\sigma^2(1 - \tau_t/\alpha)\,\mathrm{d}t.$$

Duality with respect to this deterministic process implies that the filter evolves within the Gaussian family and the computational cost is linear in n, that is, we are dealing with a finite-dimensional filter.

Of course, all this is known: the ODEs above are the well-known equations for the first two moments of linear SDEs, and the filter is the Kalman filter. Thus, within the assumptions we have made in this article, the finite-dimensional filter corresponds to the special case where the dual is purely deterministic. We considered $d = 1$ for simplicity, but the results carry over to multi-dimensional stationary linear SDEs. The same discussion as for the CIR applies here regarding the restrictions posed by needing that $h \in \mathscr{D}(\mathcal{A})$. We return to this issue in Section 4.

3.3. Wright–Fisher diffusions

Wright–Fisher (WF) processes are K-dimensional diffusions with paths confined in the $(K-1)$-dimensional simplex

$$\Delta_K = \left\{ x \in [0,1]^K : \sum_{i=1}^{K} x_i = 1 \right\}. \tag{23}$$

These processes approximate, among others, large-population discrete Wright–Fisher reproductive models with non-overlapping generations, and describe the time-evolution of the species abundancies when the individuals in the underlying population are subject to random genetic drift and, possibly, mutation, selection and recombination. See, for example, Chapter 5 in [7] and Chapter 10 in [10]. Here we are interested in the case without selection nor recombination, and with *parent-independent mutation*. That is, we consider a WF diffusion with generator

$$\mathcal{A} = \frac{1}{2} \sum_{i=1}^{K} \big(\alpha_i - |\boldsymbol{\alpha}| x_j\big) \frac{\partial}{\partial x_i} + \frac{1}{2} \sum_{i,j=1}^{K} x_i (\delta_{ij} - x_j) \frac{\partial^2}{\partial x_i \, \partial x_j}, \tag{24}$$

where δ_{ij} denotes the Kronecker delta, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}_+^K$ and $|\boldsymbol{\alpha}| = \sum_{i=1}^K \alpha_i$. The domain of the operator $\mathcal{A}$ is taken to be $C^2(\Delta_K)$, and the closure of $\mathcal{A}$ generates a strongly continuous contractive semigroup on $C(\Delta_K)$. See [9] for details. Note that this is a hypoelliptic diffusion, that is, the square of the diffusion matrix is not full rank, as a result of the constraint $\sum_i x_i = 1$. Even though we could work with an elliptic diffusion for the $K-1$ variables, it is the formulation above that is desirable for identifying the dual, as we will show.

Such diffusion is reversible with respect to the Dirichlet distribution

$$\pi(\mathrm{d}x_1, \ldots, \mathrm{d}x_K) = \frac{\Gamma(|\boldsymbol{\alpha}|)}{\prod_{j=1}^K \Gamma(\alpha_j)} x_1^{\alpha_1 - 1} \cdots x_K^{\alpha_K - 1} \, \mathrm{d}x_1 \cdots \mathrm{d}x_K, \qquad x \in \Delta_K. \tag{25}$$

In this model, we have $d = K \geq 2$ and $l = 0$, therefore there is no deterministic component in the dual process. We denote

$$x^{\mathbf{m}} = x_1^{m_1} \cdots x_K^{m_K}, \qquad x \in \Delta_K, \mathbf{m} \in \mathcal{M},$$

and define

$$h(x,\mathbf{m}) = \frac{\Gamma(|\boldsymbol{\alpha}|+|\mathbf{m}|)}{\Gamma(|\boldsymbol{\alpha}|)}\prod_{j=1}^{K}\frac{\Gamma(\alpha_j)}{\Gamma(\alpha_j+m_j)}x^{\mathbf{m}}, \tag{26}$$

whence clearly $h(\cdot,\mathbf{m}) \in \mathscr{D}(\mathcal{A})$. This can be identified with the Radon–Nikodym derivative between a Dirichlet distribution with parameters $(\alpha_1+m_1,\dots,\alpha_K+m_K)$ and π, and it is conjugate to emission densities that as a function of x are proportional to

$$x_1^{n_1}\cdots x_K^{n_K}, \qquad n_i \in \mathbb{Z}_+, i=1,\dots,K,$$

in which case t in A2 coincides with $t(y,m) = n+m$. Such type of emission density arises, for example, for observations $Y_n = (n_1,\dots,n_K)$ distributed as Multinomial with parameters $X_{t_n} = (X_{t_n,1},\dots,X_{t_n,K})$, giving rise to a dynamic version of the Dirichlet-Multinomial conjugate Bayesian model.

Then we have

$$\begin{aligned}\mathcal{A}h(x,\mathbf{m}) &= \sum_{i=1}^{K}\left(\frac{\alpha_i m_i}{2}+\binom{m_i}{2}\right)\frac{\Gamma(|\boldsymbol{\alpha}|+|\mathbf{m}|)}{\Gamma(|\boldsymbol{\alpha}|)}\prod_{j=1}^{K}\frac{\Gamma(\alpha_j)}{\Gamma(\alpha_j+m_j)}x^{\mathbf{m}-\mathbf{e}_i}\\ &\quad -\sum_{i=1}^{K}\left(\frac{|\boldsymbol{\alpha}|m_i}{2}+\binom{m_i}{2}+\frac{1}{2}m_i\sum_{j\neq i}m_j\right)\frac{\Gamma(|\boldsymbol{\alpha}|+|\mathbf{m}|)}{\Gamma(|\boldsymbol{\alpha}|)}\prod_{j=1}^{K}\frac{\Gamma(\alpha_j)}{\Gamma(\alpha_j+m_j)}x^{\mathbf{m}}\\ &= \frac{|\boldsymbol{\alpha}|+|\mathbf{m}|-1}{2}\sum_{i=1}^{K}m_i h(x,\mathbf{m}-\mathbf{e}_i) - \frac{|\mathbf{m}|(|\boldsymbol{\alpha}|+|\mathbf{m}|-1)}{2}h(x,\mathbf{m}).\end{aligned}$$

This suggests considering a one-component dual process, with M_t a Markov jump process with generator A obtained by letting

$$\lambda\big(|\mathbf{m}|\big) = \big(|\boldsymbol{\alpha}|+|\mathbf{m}|-1\big)/2, \qquad \rho(\theta)\equiv 1,$$

in (16). Since $h(x,\cdot)\in\mathscr{D}(A)$, it is then easy to check that the local duality condition

$$\mathcal{A}h(\cdot,\mathbf{m})(x) = Ah(x,\cdot)(\mathbf{m})$$

holds. Hence, the WF diffusion with parent-independent mutation X_t and the death process M_t on $\mathbb{Z}_+^K$, which jumps from $\mathbf{m}$ to $\mathbf{m}-\mathbf{e}_j$ at rate $m_j(|\boldsymbol{\alpha}|+|\mathbf{m}|-1)/2$, are dual with respect to the above h in the sense of A3. The transition probabilities of M_t are as in Proposition 2.1.

Filtering the WF model when $K=2$ on the basis of binomial data was studied in [5]. One can appreciate the strength of the approach we introduce here, since it is straightforward to obtain the filtering recursion using the dual and Proposition 2.1 for any K. It has to be noted that, in our opinion, one of the reasons why the results are harder to obtain using the approach in [5], is because they decide to work with the elliptic WF model, which is a scalar diffusion since $K=2$. Working with the elliptic model hides the structure of duality, which is immediately apparent in the hypoelliptic model.

The death process we obtain in this section can be seen as a special case of the process used in [1] for deriving an infinite mixture expansion for the transition kernel of the WF diffusion with selection.

An extension of WF diffusions to the case of infinitely-many types is given by Fleming–Viot processes. These are measure-valued diffusions whose finite-dimensional projections onto partitions of the type space coincide with WF processes. A duality relation holds between the Fleming–Viot process and a function-valued process related to Kingman's coalescent. See, for example, [11]. However, by applying to such dual process the same finite-dimensional projection that yields the WF process, one does not obtain the dual derived here, since binning the process into finitely-many sets hides some important information about the events at the level of particles.

4. Discussion

We have demonstrated that computable filtering follows from duality, in the sense described in Assumptions A2 and A3 in Section 2. A sufficient condition to establish duality is the local duality described in Assumption A4, which is based on the properties of the generator of the signal process and its relation to the semigroup operator via the Kolmogorov backward equation. The use of this functional analytic machinery places some constraints on the duality function in A2, such as for example that as a function of x it has to vanish at infinity. Therefore, even when duality holds in the sense of A3 for functions that do not satisfy such constraints, the local duality cannot be used to prove this. On the other hand, the local duality can still be used formally to identify the dual. Both in the CIR process when $\theta < \gamma/\sigma^2$ and in the OU process when $\tau < \alpha$ (see Section 3 for details) the formal application of the generator identifies the dual correctly.

We have assumed reversibility with respect to a probability measure π. In fact, our methodology relies on the existence of such reversible measure but does not require that it be a probability measure. An inspection of Proposition 2.2 reveals that it is still valid in this more general case, provided $h(x, \mathbf{m}, \theta)\pi(\mathrm{d}x)$ is a probability measure. Therefore, h is the Radon–Nikodym derivative between the measures in $\mathcal{F}$ and the reversible measure, in this more general framework which also covers non-stationary signals.

Another topic of investigation is the connection of the duality, as used in this paper, and results about the spectral representation of the transition kernel of the signal, for example, the type of expression in (22). There are classic results about such expressions, see, for example, Chapter 13 of [18], and their existence seems to be related to computable filtering, see Section 6.4 of [5], but the connection is not well understood.

Our results in Section 2 show that for observations generated by a stationary process the computational cost associated with the identification of the filtering distributions grows polynomially with the number of observations, unless $K = 0$ in which case the growth is linear. However, it might be the case that most of the components in the mixture representations have negligible weight. Previous simulation studies, see, for example, Table 1 in [13], show that after a few iterations the filter might concentrate all its mass in two or three components. We believe that the connection to the dual process might be very helpful in studying the effective number of components. However, there are subtleties in this line of research. Note that when X_t is ergodic, and $t_i - t_{i-1}$ is large relative to its mixing time, practically all mass of the filtering distribution will be

concentrated on a single component, the "root" $(\mathbf{0}, \tilde{\theta})$ (see A2) that corresponds to the invariant measure π. Therefore, the time evolution of the number of states with non-negligible filtering probabilities (say above a given $\varepsilon \approx 0$) will depend on the number of observations per unit of time in the X process. This aspect deserves careful study.

We have obtained explicit filters also for a class of Fleming–Viot diffusions with parent independent mutation and for a class of measure-valued branching processes with immigration, under parametrizations which make them reversible with respect to the Dirichlet and the gamma process respectively. A peculiarity of this framework, entailed by the fact that the signal is measure-valued and the observations are random draws from the signal, is the lack of a common dominating measure for the emission distributions, hence the lack of likelihood, which makes the nature of the problem truly non-parametric. The techniques for obtaining an optimal filter thus necessarily differ from those illustrated in this paper and will be reported elsewhere.

Appendix: Proof of Proposition 2.1

Before stating the result, we recall a useful lemma, whose proof can be found in [5].

Lemma A.1.

$$\sum_{j=0}^{l} \frac{(-1)^j}{(\lambda_n - \lambda_{n-1-j}) \prod_{0\le h\le l, h\neq j} |\lambda_{n-1-j} - \lambda_{n-1-h}|} = \frac{1}{\prod_{1\le h\le l+1} (\lambda_n - \lambda_{n-h})}.$$

Proof of Proposition 2.1. Consider first the one-dimensional case, that is, $\mathbf{m} = m$, denote for brevity $\theta[s,t] = \int_s^t \rho(\theta_u)\,\mathrm{d}u$, and define for $i \ge 1$

$$I_{1,\dots,i} = \int_0^t \cdots \int_{t_{i-1}}^t \mathrm{e}^{-\lambda_m \theta[0,t_1]} \prod_{k=1}^{i-1} \theta_{t_k} \mathrm{e}^{-\lambda_{m-k}\theta[t_k,t_{k+1}]}\,\mathrm{d}t_k\, \theta_{t_i} \mathrm{e}^{-\lambda_{m-i}\theta[t_i,t]}\,\mathrm{d}t_i,$$

$$I_{1,\dots,j-1,j,\dots,i} = \int_0^t \cdots \int_{t_{i-1}}^t \mathrm{e}^{-\lambda_m \theta[0,t_1]} \prod_{k=1,k\neq j}^{i-1} \theta_{t_k} \mathrm{e}^{-\lambda_{m-k}\theta[t_k,t_{k+1}]}\,\mathrm{d}t_k\, \theta_{t_i} \mathrm{e}^{-\lambda_{m-i}\theta[t_i,t]}\,\mathrm{d}t_i,$$

where $t_j := t_{j+1}$ in $I_{1,\dots,j-1,j+1,\dots,i}$. It can be easily seen that

$$I_i = \frac{\mathrm{e}^{-\lambda_{m-i}\theta[0,t]} - \mathrm{e}^{-\lambda_m \theta[0,t]}}{\lambda_m - \lambda_{m-i}}. \tag{A.1}$$

Then we have

$$\left(\prod_{h=0}^{i-1} \lambda_{m-h}\right)^{-1} p_{m,m-i}(t) = I_{1,\dots,i}, \tag{A.2}$$

where $p_{m,m-i}(t)$ is the transition probability associated to the one-dimensional death process. By integrating twice, we obtain

$$\begin{aligned} I_{1,\dots,i} &= \frac{(-1)(I_{1,\dots,i-1} - I_{1,\dots,i-2,i})}{\lambda_{m-(i-1)} - \lambda_{m-i}} \\ &= \frac{(-1)^2}{\lambda_{m-(i-1)} - \lambda_{m-i}} \left[\frac{(I_{1,\dots,i-2} - I_{1,\dots,i-3,i-1})}{\lambda_{m-(i-2)} - \lambda_{m-(i-1)}} - \frac{(I_{1,\dots,i-2} - I_{1,\dots,i-3,i})}{\lambda_{m-(i-2)} - \lambda_{m-i}} \right]. \end{aligned}$$

The iteration of the successive integrations can be represented as a binary tree with root $(i,0) := I_{1,\dots,i}$, whose node $(i-j, i-k) := I_{1,\dots,i-j,i-k}$ branches into $(i-j,0) := I_{1,\dots,i-j}$ and $((i-j-1)^+, i-k) = I_{1,\dots,i-j-1,i-k}$, with both branches weighed $1/(\lambda_{m-(i-j)} - \lambda_{m-(i-k)})$, determined by the parent node's indices. The leaves correspond to nodes where the left coordinate touches zero if the right coordinate is already zero, or where the left crosses zero if the right coordinate is positive. The term associated to the leaf $(0, i-k)$ will be $(-1)^i \mathrm{e}^{-\lambda_{m-(i-k)}\theta[0,t]}$ weighed by some appropriate coefficient. The level before the leaves can be seen as the sequence

$$\underbrace{\underbrace{I_1 I_2}_{2^1}\ \underbrace{I_1 I_3}_{2^1}}_{2^2}\ \underbrace{\underbrace{I_1 I_2 I_1 I_4}_{2^2}\ \underbrace{I_1 I_2 I_1 I_3 I_1 I_2 I_1 I_5}_{2^3}}_{2^3} \dots,$$

where every sequence of 2^i terms is repeated with the last index augmented by one, and each I_i produces the leaves $\mathrm{e}^{-\lambda_m}$ and $\mathrm{e}^{-\lambda_{m-i}}$. Hence given i, there are 2^{i-2} terms I_1, 2^{i-3} terms $I_2, \dots,$ 2^1 terms I_{i-2}, 2^0 terms I_{i-1} and I_i. Note also that I_1 has 2^0 paths in common with I_i, 2^0 paths in common with I_{i-1}, 2^1 paths in common with $I_{i-2}, \dots, 2^{i-3}$ paths in common with I_2.

The correct coefficient for I_k is computed by collecting some constants related to the paths that have the same last coefficient and simplifying. In particular, given i, the paths to be grouped for I_k are those whose constants for indices greater than k change, since according to the rule above, when k is the rightmost index in $I_{1,\dots,k}$, there is only one path down to I_k. Hence, given i, term I_k has coefficient

$$\begin{aligned} &\frac{(-1)^{i-1}}{\prod_{1\le h<k}(\lambda_{m-h} - \lambda_{m-k})} \\ &\quad \times \sum_{j=0}^{i-(k+1)} \frac{1}{(\lambda_{m-k} - \lambda_{m-k-1-j}) \prod_{0\le h\le i-(k+1), h\ne j} (\lambda_{m-k-1-j} - \lambda_{m-k-1-h})}. \end{aligned}$$

By taking moduli and applying Lemma A.1 to the sum above, we obtain

$$\frac{(-1)^{i-1}}{\prod_{1\le h\le i, h\ne k}(\lambda_{m-k} - \lambda_{m-h})}.$$

The result now follows from (A.1) and (A.2), and from the fact that in the K-dimensional case, the probability of going from $\mathbf{m}$ to $\mathbf{m}-\mathbf{i}$, conditional on $|\mathbf{i}|$, is

$$p(i_1,\ldots,i_K;\mathbf{m},|\mathbf{i}|)=\frac{\binom{m_1}{i_1}\cdots\binom{m_K}{i_K}}{\binom{|\mathbf{m}|}{|\mathbf{i}|}}.$$

□

Acknowledgements

The second author is supported by the European Research Council (ERC) through StG "N-BNP" 306406. The authors would like to thank Valentine Genon-Catalot and Aleksandar Mijatović for helpful discussions.

References

[1] Barbour, A.D., Ethier, S.N. and Griffiths, R.C. (2000). A transition function expansion for a diffusion model with selection. *Ann. Appl. Probab.* **10** 123–162. MR1765206

[2] Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Chichester: Wiley. MR1274699

[3] Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in Hidden Markov Models*. *Springer Series in Statistics*. New York: Springer. With Randal Douc's contributions to Chapter 9 and Christian P. Robert's to Chapters 6, 7 and 13, with Chapter 14 by Gersende Fort, Philippe Soulier and Moulines, and Chapter 15 by Stéphane Boucheron and Elisabeth Gassiat. MR2159833

[4] Chaleyat-Maurel, M. and Genon-Catalot, V. (2006). Computable infinite-dimensional filters with applications to discretized diffusion processes. *Stochastic Process. Appl.* **116** 1447–1467. MR2260743

[5] Chaleyat-Maurel, M. and Genon-Catalot, V. (2009). Filtering the Wright–Fisher diffusion. *ESAIM Probab. Stat.* **13** 197–217. MR2518546

[6] Cox, J.C., Ingersoll, J.E. Jr. and Ross, S.A. (1985). A theory of the term structure of interest rates. *Econometrica* **53** 385–407. MR0785475

[7] Dawson, D.A. (1993). Measure-valued Markov processes. In *École d'Été de Probabilités de Saint-Flour XXI—*1991. *Lecture Notes in Math.* **1541** 1–260. Berlin: Springer. MR1242575

[8] Etheridge, A.M. (2000). *An Introduction to Superprocesses*. *University Lecture Series* **20**. Providence, RI: Amer. Math. Soc. MR1779100

[9] Ethier, S.N. and Kurtz, T.G. (1981). The infinitely-many-neutral-alleles diffusion model. *Adv. in Appl. Probab.* **13** 429–452. MR0615945

[10] Ethier, S.N. and Kurtz, T.G. (1986). *Markov Processes*. *Characterization and Convergence*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. New York: Wiley. MR0838085

[11] Ethier, S.N. and Kurtz, T.G. (1993). Fleming–Viot processes in population genetics. *SIAM J. Control Optim.* **31** 345–386. MR1205982

[12] Feller, W. (1951). Diffusion processes in genetics. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 1950 227–246. Berkeley and Los Angeles: Univ. California Press. MR0046022

[13] Genon-Catalot, V. and Kessler, M. (2004). Random scale perturbation of an AR(1) process and its properties as a nonlinear explicit filter. *Bernoulli* **10** 701–720. MR2076070

[14] Griffiths, R.C. (2006). Coalescent lineage distributions. *Adv. in Appl. Probab.* **38** 405–429. MR2264950

[15] Hutzenthaler, M. and Wakolbinger, A. (2007). Ergodic behavior of locally regulated branching populations. *Ann. Appl. Probab.* **17** 474–501. MR2308333

[16] Jansen, S. and Kurt, N. (2013). On the notion(s) of duality for Markov processes. Available at arXiv:1210.7193 [math.PR].

[17] Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions. Vol.* 1, 2nd ed. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. New York: Wiley. MR1299979

[18] Karlin, S. and Taylor, H.M. (1981). *A Second Course in Stochastic Processes*. New York: Academic Press. MR0611513

[19] Kawazu, K. and Watanabe, S. (1971). Branching processes with immigration and related limit theorems. *Theory Probab. Appl.* **16** 34–51.

[20] Lévy, P. (1948). *Processus Stochastiques et Mouvement Brownien*. Paris: Gauthier-Villars & Cie.

[21] Liggett, T.M. (2005). *Interacting Particle Systems. Classics in Mathematics*. Berlin: Springer. Reprint of the 1985 original. MR2108619

[22] Sen, A. and Balakrishnan, N. (1999). Convolution of geometrics and a reliability problem. *Statist. Probab. Lett.* **43** 421–426. MR1707953

[23] Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoret. Population Biol.* **26** 119–164. MR0770050

Received May 2013